Multiword Expressions in Spoken Language: an exploratory study on pronunciation variation

Diana Binnenpoorte [a] Catia Cucchiarini [a] Lou Boves [a]
Helmer Strik [a]

[a] *Radboud University Nijmegen, The Netherlands*

Abstract

The study presented in this paper was aimed at exploring the possibilities of modelling specific pronunciation characteristics of multiword expressions (MWEs) for both automatic speech recognition (ASR) and automatic phonetic transcription (APT). For this purpose we first drew up an inventory of frequently found N-grams extracted from orthographic transcriptions of spontaneous speech contained in a large corpus of spoken Dutch. These N-grams were filtered and subsequently assigned to linguistic categories. For a small selection of these N-grams we examined the phonetic transcriptions contained in the corpus. We found that the pronunciation of these N-grams differed to a large extent from the canonical form. In order to determine whether this is a general characteristic of spontaneous speech or rather the effect of the specific status of these N-grams, we analysed the pronunciations of the individual words composing the N-grams in two context conditions: 1) in the N-gram context and 2) in any other context. We found that words in N-grams do indeed have peculiar pronunciation patterns. This seems to suggest that the N-grams investigated may be considered as MWEs that should be treated as lexical entries in the pronunciation lexicons used in ASR and APT, with their own specific pronunciation variants.

Key words: multiword expressions; automatic phonetic transcription; automatic speech recognition; spontaneous speech; pronunciation variation

1 Introduction

Multiword expressions (MWEs) have been studied in theoretical linguistics (Nunberg et al., 1994; Sag et al., 2001; Wong-Fillmore, 1979), and more recently also in NLP (Koster, 2004; Nivre and Nilsson, 2004; Odijk, 2004). So far,

most of the research on MWEs has concerned their extraction and handling in written language. However, it has also long been known that frequently used sequences of words, whether they are stock phrases (e.g. *I don't know*) or lexicalized idiomatic expressions (e.g. *kick the bucket*), show pronunciation phenomena that have not been observed when the words occur in less frequent contexts (cf. the pronunciations of '*I don't know*' in Hawkins (2003)). While observations such as Hawkins' are to some extent anecdotal, the advent of large spoken language corpora has made it possible to investigate pronunciation variation in multiword expressions quantitatively. In this paper we investigate pronunciation variation in MWEs in a large corpus of spontaneously spoken Dutch (Oostdijk, 2002). Although the Spoken Dutch Corpus (also known as CGN) also comprises more formal speech styles, we focus on spontaneous speech because we think that the problem of pronunciation variation in MWEs is most acute in this style. Speech recognition performance for spontaneous speech is way below the performance for read speech (Pallett, 2003) and there are indications that a large proportion of the performance gap is due to the inability to model pronunciation variation in spontaneous speech effectively (Strik and Cucchiarini, 1999).

For ASR it has been found that simply adding the most frequent pronunciation variants of individual words to the lexicon becomes counter-productive as soon as the average number of variants per word exceeds a threshold of about 2.5 (Kessens et al. 2003; Yang and Martens, 2000). At the same time, it appears that adding frequent bigrams to the lexicon and treating these as words with their own specific pronunciation variants does improve ASR performance (Beulen et al., 1998; Finke and Waibel, 1997; Kessens et al., 1999; Sloboda and Waibel, 1996). However, in these studies the notion of MWE is mainly deployed for the benefit of reducing word error rate in ASR. No special attention was given to the lexical and linguistic role and status of the word sequences. In the present paper we investigate whether it is indeed true that words in MWEs in spontaneous speech have more -and specifically more reduced- pronunciation variants than when the same words occur in a general context.

In our research we first extracted frequent word sequences (which we will call MWEs for convenience throughout this paper) from all spontaneous speech recordings in the Spoken Dutch Corpus (CGN), which we then analyzed to determine their lexical status and syntactic structures. Then we proceeded to a more detailed analysis of MWEs in that part of the CGN that comes with manually verified broad phonetic transcriptions. In doing so, we focused on reduction phenomena, and we tried to determine whether there is a relation between the degree of reduction in a given MWE and the lexical/syntactic category to which it belongs.

2 MWEs in the Spoken Dutch Corpus

MWEs were extracted from the Spoken Dutch Corpus, a database containing about 9 million words of contemporary Dutch as spoken in the Netherlands and Flanders. All recordings are orthographically transcribed, lemmatised and enriched with part-of-speech (POS) information. For about 900,000 words, more detailed annotations are available, such as a manual broad phonetic transcription, a hand-checked word alignment, syntactic annotation and prosodic information. This sub-corpus of 900,000 words, called the *core corpus*, was composed in such a way that it faithfully reflects the design of the full corpus (Oostdijk, 2002). The speech material in the corpus was recorded in various socio-situational settings from speakers of different age, sex, educational level and region of birth. The speech material collected consists of various speech styles, varying from read speech recorded in a studio environment with professional speakers, through interviews which are more or less prepared dialogues, and business negotiations to spontaneous dialogues recorded in home environments.

For our study we are only interested in spontaneous speech; therefore, only speech styles that can be characterized as spontaneous or extemporaneous were selected. In order to make a comprehensive inventory of MWEs in unprepared speech, we used the orthographic transcriptions of all lessons (LS), spontaneous dialogues (SD), and spontaneous telephone conversations (ST). The conversational settings differ among the three components. In the LS component a teacher discusses and explains several subjects with a group of students. In the SD component two or more people have a face-to-face conversation in a home environment, often about objects in the room or activities such as game playing that they are involved in. Finally, in the ST component two friends or family members have a telephone conversation without the need to talk about specific topics. Table 1 summarizes the characteristics of the material that are most important for the present study.

Table 1
Total duration of the components, number of words and number of different speakers involved.

speech style	duration (hh:mm:ss)	# words	# speakers
LS	30:41:04	299,973	398
SD	149:44:17	1,747,789	231
ST	92:24:50	1,253,741	534
total	272:50:11	3,301,503	1,148

2.1 Criteria for selecting N-grams as MWEs

There is no generally accepted definition of the concept of MWEs in spoken language. Therefore we based our investigations on what we consider a reasonable operational definition of the concept, adapted to the specific requirements of our study. Since we are interested in the effect of MWE status on pronunciation variation, our first criterion was that only contiguous sequences of words qualify. We expect to see substantial pronunciation variation in the form of cross-word assimilation and degemination. In lexicalized MWEs that are broken by interspersed words, the cross-word phonetic context of the contiguous MWE no longer exists. Consequently, one cannot expect to observe the cross-word assimilations and reductions that may be characteristic for the contiguous MWEs. A practical advantage of this criterion is that it allows us to start the search for potential MWEs by simply creating lists of sequences of N words with a frequency of occurrence that is higher than what one would expect for arbitrary syntactically correct sequences.

Thus, we started the search for N-grams that might qualify as MWE by extracting all 3-, 4-, 5,- and 6-grams from the orthographic transcription files. In doing so, we used the -admittedly somewhat arbitrary- criterion proposed in chapter 13 in Biber et al. (1999) to establish the minimum frequency that a sequence should exceed in order to qualify as '*exceptionally frequent*'. Expressions containing three or four words should have a minimal frequency of 10 per million words, and expressions containing more than four words should have 5 or more occurrences per million words. In our case, with a source text of 3.3M words, we require the frequency of a unique 3-gram and 4-gram to be at least 30, and for the 5-gram and 6-gram at least 15.

Because we want to use frequent sequences to investigate pronunciation variation in word sequences that may qualify as MWEs, or at least as stock phrases, we decided to apply a number of additional criteria to filter the raw lists of expressions that exceed Biber's frequency threshold. First, we did not want to include word sequences that straddle a deep syntactic boundary. These are likely to induce pauses between the words on either side of the boundary that block assimilation and degemination processes. The only clues for syntactic boundaries in the CGN transcriptions are full stops, question marks, and ellipsis marks; no commas and other 'minor' punctuation marks are included. Therefore, we restricted the search for MWEs to sequences that do not include one of the three punctuation marks.

A second criterion in the filter process was the length of the sequences. Given the size of the corpus, we did not expect to find frequent sequences longer than six words. For theoretical and practical reasons we decided to omit bigrams. For one thing, many frequent bigrams are part of frequent N-grams with $N > 2$, so that we can observe and analyze their pronunciation variation even if we do not include bigrams. Moreover, the number of frequent bigrams is extremely large, and the sheer number complicates analysis considerably.

Therefore, we decided to take $3 \leq N \leq 6$.
Third, we decided to exclude disfluencies and hesitations from our corpus of frequent N-grams. The initial N-gram list contained a substantial number of frequent sequences in which one or more filled pause markers were present. In the CGN all filled pauses are transcribed by one of two 'hesitation' words, '*uh*' and '*uhm*'. This transcription convention is part of the explanation why word sequences containing filled pause markers occurred so frequently. Another part of the explanation is definitively related to the fact that filled pauses and hesitations do not occur in random positions, but tend to occur just before content words, due to which sequences such as '*in the uhm*' are rather frequent. Although detecting and handling hesitations and disfluencies is of crucial importance for automatic recognition of spontaneous speech, we feel that these phenomena form a research topic in their own right, probably related, but also somewhat independent of pronunciation variation in MWEs. Therefore, we excluded N-grams such as '*de uh de uhm*' ('*the eh the ehr*') as potential MWEs. Sequences containing '*ggg*' (the symbol for speaker noise) or '*xxx*' (unintelligible speech) were excluded for the same reason.
Fourth, we also decided to exclude repetitions. In the spontaneous part of the CGN one can distinguish two different categories of repetitions. The first category, which comprises sequences such as '*en de en de*' ('*and the and the*'), represents what are likely to be disfluencies. These cases are rejected for the reason explained above. The second category is perhaps more problematic. It contains sequences such as '*ja ja ja ja*' ('*yes yes yes yes*'), which may be related to disfluencies, but which can also be used to indicate emphasis or other pragmatic effects. The CGN transcriptions do not provide information that can be used to distinguish disfluencies from truly linguistic devices, such as for lending emphasis or expressing sarcasm. For this reason we decided to remove all two and three word repetitions from the lists of possible MWEs.
The last criterion that we used to filter the lists of frequent N-grams is the requirement that the sequence should have higher than expected frequency in all three sub-corpora (LS, SD, ST). This stipulation removes sequences such as '*een twee drie vier*' ('*one two three four*'), which are frequent in the SD sub-corpus, due to the fact that the speakers were encouraged to play games to keep the conversation going. Perhaps it might be possible to identify and eliminate setting-specific sequences on the basis of linguistically informed rules, but it is very difficult to formulate adequate rules. Thus, we used the uniform presence criterion to detect and remove such artefacts from the lists. Table 2 summarizes the results of the MWE extraction on the 3.3M word spontaneous speech part of the CGN. It can be seen that both the number of types and the token/type ratio decrease as the sequences grow longer. The number of types would have been much larger if we had not applied the criterion that expressions should occur with higher than expected frequency in all three sub-corpora. That criterion removed many sequences from the subcorpus of face-to-face dialogs that were directly related to playing card or board games. Removing setting specific types resulted in a large increase in

Table 2
Number of types and tokens of N-grams passing the selection criteria.

	3-grams	4-grams	5-grams	6-grams
# types	3,015	247	48	1
# tokens	217,230	13,495	1,285	19
token/type ratio	72.05	54.63	26.71	19

the average token/type ratio.
From Table 2 it can be deduced that the 3,311 N-gram types cover about 21% of the source corpus. Apparently spontaneous conversations consist to a large extent of 'stock phrases' and/or true MWEs. As not many generalisations can be made over one type, the one remaining 6-gram will not be considered in the remainder of the paper.

2.2 Categorization of selected N-grams

Once the MWEs had been extracted from the transcription files, we proceeded to classify them manually into six broad categories:

(1) The N-gram constitutes a whole grammatical sentence.
E.g. *'weet ik veel' (I've no idea)*
(2) The N-gram constitutes a grammatical constituent.
E.g. *'op een andere manier' (in a different way)*
(3) The N-gram constitutes an interjection.
E.g. *'nou ja goed' (well alright)*
(4) The N-gram constitutes the beginning of a possible main clause.
E.g. *'en dan moet je' (and then you have to)*
(5) The N-gram constitutes the beginning of a possible subordinate clause.
E.g. *'als het goed is' (if it is okay)*
(6) The N-gram cannot be classified in any of the above and is categorized as 'other'.
E.g. *'weet niet of je' (don't know whether you)*

These categories emerged during the process, based on our interpretation of the MWEs. The categories fall apart in two broad classes; the first three categories include complete syntactic units, whereas the last three include sequences of words that do not constitute a complete syntactic unit. The distribution of the categories of the MWE types is displayed in Table 3.
Although the classification results in Table 3 are instructive, it should be noted that many MWEs assigned to the categories 2 to 5 would be moved to another class if some highly frequent function word were added before or after the sequence. Thus, the classification is to some extent based on evidence that is not extremely reliable. It would be worthwhile to repeat the experiment with

Table 3
Distribution of categories expressed in number and percentage.

	3-gram	%	4-gram	%	5-gram	%
1. complete sentence	163	5.4	25	10.1	9	18.7
2. constituent	260	8.6	18	7.3	3	6.3
3. interjection	64	2.1	12	4.9	5	10.4
4. begin of main clause	1002	33.2	124	50.2	22	45.8
5. begin of subordinate clause	126	4.2	4	1.6	0	0.0
6. other	1537	51.0	71	28.7	14	29.2
total	3152	104.5	254	102.8	53	110.4
categorized twice	137	4.5	7	2.8	5	10.4
# types	3015	100.0	247	100.0	48	100.0

a mix of words and POS information, and count the frequency of sequences of the form $POS_x, word_1, \cdots, word_n$ and $word_1, \cdots, word_n, POS_y$, where POS_x indicates a set of words with the POS-tag x. Some trends emerge from this table. In general, for all three N-gram types, the contribution of N-grams classified as incomplete syntactic units (category 4, 5, and 6) is much larger than the contribution of those classified as complete syntactic units. During the selection procedure no restrictions on syntactic completeness were applied, because syntax annotation is only available for the *core corpus* in the CGN. Moreover, in Kessens et al. (1999) it is shown that modelling pronunciation variation of highly frequent sequences of words does improve recognition performance, but these word sequences need not constitute syntactic units.
The majority of the N-grams belong to category 4, where the N-gram constitutes the beginning of what is likely to become a main clause. In Dutch given information tends to go to the beginning of a clause, whereas new information tends to occur at the end. The high proportion of conventional expressions at the beginning of a clause may well help speakers to overlap cognitive processing needed to express the new information with almost automatic generation of the beginning of the sentence or clause in which the new information is embedded. Listeners may also profit from such an alternation of predictable and new information. In any case, the high frequency of a small number of clause-initial 'formulae' suggests that in conversational Dutch the variety of introductory clauses is not very broad. This impression is corroborated by the fact that the average number of tokens per type in the N-grams in category 4 is relatively high. Therefore, the frequently used N-grams at the start of a main clause actually occur more often than might appear from the figures in Table 3, which only refer to types.
In the collection of the 3-grams the proportion of the 'other' category is larger than that of 'begin of main clause'. This might indicate that a sequence length of three words is too short to be identified as a possible beginning of a main clause or even a syntactic unit. This hypothesis is in line with the observation that adding one word at the beginning or end of a sequence often would change its category assignment. Alternatively, conversational speech may contain a substantial number of frequent word sequences that straddle the boundary

between NP, PP or AP [1] constituents. Future research, in which POS (and perhaps also syntactic annotation) is used will show which possibility is more likely.
When the length of the sequences increases, the share of complete sentences and multiword interjections (category 1 and 3) also increases. The prominent presence of long interjections motivated the creation of category 3, as a special case of category 2 during the course of the classification process. In this context it is interesting to observe that the proportion of complete grammatical constituents which are not a sentence or an interjection decreases when the sequence length grows. This may indicate that highly frequent constituents (NPs, PPs and APs) mainly consist of three words in conversational Dutch.

3 Pronunciation variation in MWEs

Having compiled the lists of MWEs and some data on the occurrences extracted from the spontaneous speech in the CGN, we proceed to investigate whether words in MWEs have more reduced pronunciation variants than when the same words occur in another arbitrary context. This part of the study is limited by necessity to the '*core corpus*' in CGN, i.e., the part that comes with manually verified broad phonetic transcriptions. On average, the *core corpus* covers 10% of the total corpus. In Table 4 the size and other characteristics of the spontaneous components of the *core corpus* are displayed. From a comparison with the figures in Table 1 it can be seen that the spontaneous speech styles are represented proportionally in the *core corpus*.

Table 4
Duration, number of words and number of different speakers in the spontaneous components of the *core corpus*.

Speech style	duration (hh:mm:ss)	# words	# speakers
LS	2:43:36	25,961	48
SD	9:43:39	106,182	108
ST	14:42:28	201,141	101
Total	27:09:43	333,284	255

3.1 Selection of frequent N-grams for pronunciation analysis

The analysis of the effect of the frequency of N-grams on pronunciation variation can only be performed on those N-grams that occur sufficiently frequently

[1] noun phrase, prepositional phrase, and adjective phrase respectively

to allow us to distinguish systematic from coincidental observations. This issue is all the more urgent since we now must work with a corpus of no more than 0.3 M words. There are no formal criteria to determine what 'sufficiently frequent for the purpose of analyzing pronunciation variation' is. However, it is clear that we need an absolute lower bound, in addition to the relative lower bound proposed in Biber et al. (1999) for other types of linguistic analyses. To start the analysis we decided to restrict our corpus to types which occur at least 7 times. We considered this as the minimum number that should allow at least some conclusions about the characteristics of pronunciation variants. In the 0.3M word corpus of manually transcribed spontaneous speech there were no 5- or 6-grams that fulfilled this minimum frequency criterion. Consequently, the remainder of this paper is limited to an analysis of 3-grams and 4-grams. In Table 5 the number of different N-grams for which at least 7 observations were found is displayed for the 3-grams and 4-grams, together with the mean frequency and the frequency range.

Table 5
Properties of remaining N-grams.

	3-gram	4-gram
# types	110	21
mean frequency	17.5	13.8
frequency range	7 – 118	7 – 50

We can now proceed to making an inventory of the pronunciation variants of the words that occur in frequent N-grams. The *core corpus* provides word segmentations, which connect the speech to the orthographic and phonetic transcription on the word level. This allows us to determine an unambiguous phonetic transcription for each word in the orthographic transcription.

3.2 Method of pronunciation analysis

Before we can proceed to the results of our analysis of pronunciation variation, we must first deal with two further methodological issues, viz. the way in which we defined the reference material to which we compared the pronunciation variants observed in frequent N-grams and the measure used to express differences in pronunciation variation.

3.2.1 Selection of reference material

To determine whether words occurring in frequent N-grams indeed have pronunciation variants that are different from the variants that can be observed for the same word in arbitrary but comparable contexts, we have to define the

very concept *arbitrary but comparable context*. Ideally, one would like to compare words in the same syntactic and prosodic context, only now surrounded by other words that do not form a frequent N-gram. However, since the CGN *core corpus* does not provide sufficient prosodic and syntactic information, we decided to settle for a less ambitious definition. For each word we performed an N-gram search with the restriction that only N-grams were allowed in which that specific word was in exactly the same position as in the original N-gram and that the other words in the N-gram were different from those in the original N-gram. For instance, assuming that the word '*als*' as found in the 3-gram '*als het ware*' ('*as it were*') is subject to this detailed analysis (because the 3-gram '*als het ware*' is one of the highly frequent N-grams) then only those versions of '*als*' are taken into consideration in which the two words following '*als*' do not equal '*het*' and '*ware*'.

3.2.2 Comparing different transcriptions

In order to compare the degree of discrepancy found in the conditions "only within MWE context" and "in all other contexts" (indicated as "MWE context" and "other context", respectively, in the remainder of the paper) we used the canonical transcription of each word as a reference point. More specifically, we compared the transcription of the words in the N-gram context to their canonical transcription, and we did the same with the occurrences of the words in arbitrary contexts. In this way we were able to calculate the weighted average percentage of difference for each word in the two conditions, where the weighting is based on the length of the word in question (number of segments in canonical transcription).
The differences between actually observed pronunciations and canonical representations was determined by the computer program `Align` (Cucchiarini, 1996). Table 6 shows the orthographic and canonical phonemic representations of the 4-gram *'aan de andere kant' (on the other hand)*, together with an arbitrary selection of two alternatives of the rich variety of pronunciation variants that are present in the corpus.

Table 6
Example of different pronunciations.

Orthography	*aan de andere kant*
Canonical transcription	`an d@ And@r@ kAnt`
Actual pronunciation 1	`an d Andr@ kAn`
Actual pronunciation 2	`An d And@ kAnt`

`Align` uses a dynamic programming procedure to align two sequences of phonetic symbols. It computes two kinds of distance measures, one based on an articulatory feature representation of the transcription symbols, and one based on the number of substitutions, deletions and insertions observed between the two strings in question. During the alignment procedure, proper penalties for symbol substitutions are calculated in terms of articulatory features, such as

place and manner of articulation, voice, lip rounding, length, etc. For deletions and insertions a fixed penalty is used. In addition to the feature based phonetic distance, `Align` also outputs a distance measure in the form of the percentage disagreement between the two sequences of symbols aligned. Percentage disagreement is the total number of differences between the two strings, divided by the number of segments in the canonical transcription.

$$\%disagreement = \frac{\#S + \#D + \#I}{\#phonemes} * 100\%$$

Although percentage disagreement might seem to be much coarser a measure than the feature based phonetic distance, we decided to use percentage disagreement in this study. The most important reason for doing so is that we expected that the bulk of the differences between canonical and observed pronunciations would consist of deletions in the observed pronunciations. All deletions obtain the same weight in the present version of `Align`. Moreover, results based on percentage disagreement would be easier to compare and replicate by other research teams.

3.3 Results

In the following sections we present the data concerning the actual pronunciation of the words contained in the N-grams. In Section 3.3.1 we show how these pronunciations differ from their canonical representations. Next, in Section 3.3.2 we explain and motivate a further reduction of the set of N-grams under analysis for the more detailed comparison of pronunciation variants between words in what may be MWEs and the same words occurring in arbitrary contexts, and we present the quantitative results. Finally, the results of qualitative analyses of these pronunciations are presented in Section 3.3.3.

3.3.1 N-gram pronunciation versus canonical

All the observed pronunciations of the 3-grams and 4-grams in Table 5 were aligned with the canonical representation of that specific N-gram. In the canonical representation no pronunciation variation due to context (cross-word processes) is modelled; only obligatory word internal phonological rules are applied. Although pronunciation variation due to cross-word context is very common in real speech, we choose to use this strict canonical transcription as reference material, because it is the only objective reference that can be used to generalize over contexts.
The discrepancy between the observed pronunciation and the canonical representation is expressed in percentage of substitutions, deletions and insertions

relative to the number of phonemes in the canonical representations. In Table 7 the results of the alignment of all 3-grams and 4-grams are presented, separately for the six categories from Table 3 and expressed in an average percentage of disagreement (column '%total', subdivided into substitutions, deletions and insertions) together with the number of types belonging to each category. A detailed results table for each N-gram separately can be found in Appendix 1 and 2 in Binnenpoorte (2004b). From Table 7 it can be seen that for all the 3- grams and 4-grams most of the differences between the canonical representation and the actual pronunciation are caused by deletions and substitutions of segments in the actual pronunciation. Only few insertions are observed. In quantitative terms this is precisely what one would expect: spontaneous speech is characterized by what could be considered as 'sloppy' pronunciation.

Table 7
Average percentage substitutions, deletions and insertions after alignment with canonical transcription.

	3-grams					4-grams				
	#types	%sub	%del	%ins	%total	#types	%sub	%del	%ins	%total
cat 1	32	13.89	9.14	0.47	23.50	9	15.37	13.79	1.54	30.70
cat 2	31	11.36	11.82	0.15	23.33	3	5.75	16.25	0.04	22.04
cat 3	4	11.46	15.21	0.70	27.36	2	20.33	8.66	0.00	28.99
cat 4	28	13.13	12.81	0.27	26.21	6	13.74	15.49	0.40	29.63
cat 5	1	6.00	10.00	0.00	16.00	1	3.57	15.00	0.00	18.57
cat 6	17	12.59	10.49	0.66	23.75	0	-	-	-	-

The dynamic programming algorithm used for alignment provides information not only on the number of discrepancies, but also on their nature. We found that the majority of phonemes that are deleted in the actual pronunciation of the N-grams are word final /*t*/, /*n*/ and /*r*/. Furthermore, many schwas, /@/, were deleted as well in both the 3-grams and the 4-grams. Most of the substitutions concern the reduction of full vowels in the canonical to schwas in the actual pronunciation. Many other substitutions involved the feature voice, where the unvoiced variant was most often found in the actual pronunciation. The few insertions observed seem to be related to processes that may be motivated by ease of articulation, such as homorganic glide insertion: insertion of /*j*/ or /*w*/ between two vowels (Booij, 1995), e.g. in the word '*zoiets*' (*something*). The canonical transcription is /*zoits*/, but in the observed pronunciations the most frequent form is /*zowits*/. Thus, our data form a quantitative confirmation of the abundant presence of 'sloppy speech' phenomena that have been impressionistically described for spontaneously spoken Dutch (Ernestus et al., 2002).

From Table 7 it can also be seen that the total percentage disagreement is quite similar for all the categories. Therefore, it is not possible to pursue the analysis of differences between 'true' MWEs, stock phrases and coincidental frequent word sequences in depth in the remainder of this study.

3.3.2 Effect of contexts on pronunciation of words in N-grams

Although the number of N-grams with a sufficiently high frequency in the CGN *core corpus* (cf. Table 5) does not seem impressive, it is still far too high to allow a detailed comparison of pronunciation phenomena between words in N-gram context and in arbitrary contexts. The major cause of the problem is that it is not clear whether the percentage disagreement for individual words in an N-gram can be accumulated to provide a meaningful score for the complete sequence, without thorough analysis of the phenomena that caused the discrepancies in the first place. Therefore we decided to process data manually, which requires a further reduction of the data. Because we are interested in the potential effect of MWE status on pronunciation variation, we decided to select those N-grams from the corpus summarized in Table 5 which showed the highest degree of discrepancy between the actual pronunciation and the canonical reference. In this way we selected the 10 3-grams shown in Table 8 and the 10 4-grams in Table 9.
In addition to the N-grams shown in the tables, we also had to select occurrences of all words in these N-grams in 'comparable' arbitrary contexts. As explained in section 3.2.1 we defined 'comparable arbitrary context' in terms of the position in an arbitrary N-gram, with the only additional restriction that the neighbouring words must be different from the neighbours in the MWE N-gram. The number of other contexts for a word differs enormously between the words. For example, the word '*ware*' (*were*) occurs only once outside the context '*als het ware*' (*as it were*), and the word '*een*' (*a*) from '*op een gegeven moment*' (*at a given moment*) occurs, of course, many more times.
Each individual word has two collections of pronunciations, those found in the MWE context and those found in all other contexts. The same canonical transcriptions were used as a reference for the comparison of the actual pronunciations in the two context conditions.
Comparing the percentage disagreement observed for each word in the two context conditions gives the results displayed in Tables 8 and 9. The percentage disagreement of an N-gram in one of the two contexts, is the weighted total of the average percentages disagreement of the individual words in that specific N-gram. The individual percentage disagreement of a word is normalized for the frequency of occurrence, which is different in the two contexts and varies per word. The weighting for the summation of the individual percentages disagreement is determined by the number of phonemes of the word in the reference transcription. The expressions listed in column 1 are ranked according to the difference in percentage disagreement between the two conditions. A detailed results table for each word in the N-grams can be found in Appendix 3 and 4 in Binnenpoorte (2004b).

The first observation that can be made from Tables 8 and 9 is that selecting N-grams on the basis of their pronunciation yields mainly N-grams belonging to the categories that represent complete syntactic constituents. Although

Table 8
Difference in %disagreement between two context conditions for words in 3-grams.

3-gram	category	%disagreement MWE context	%disagreement other context	difference
zoiets van ja	6	57.27	15.75	41.52
in ieder geval	2	37.17	12.26	24.91
af en toe	2	34.76	15.15	19.61
op die manier	2	31.94	12.99	18.95
't is natuurlijk	4	45.59	31.11	14.48
weet ik niet	1	29.22	21.52	7.7
dat is natuurlijk	4	34.62	28.76	5.86
hoe heet dat	1	30.43	24.95	5.48
ook helemaal niet	2	27.78	24.40	3.38
als 't ware	3	23.15	35.88	-12.73

Table 9
Difference in %disagreement between two context conditions for words in 4-grams.

4-gram	category	%disagreement MWE context	%disagreement other context	difference
dat vind 'k ook	1	48.89	29.00	19.89
op een gegeven moment	2	47.13	27.91	19.22
dat maakt niet uit	1	42.42	26.49	15.93
dat is niet zo	1 / 4	40.00	28.47	11.53
of wat dan ook	3	31.54	22.10	9.44
'k weet niet precies	4	28.57	22.73	5.84
dat weet ik niet	1	29.03	25.96	3.07
weet ik veel wat	3	26.45	25.08	1.37
dat weet ik nog	1	24.55	26.15	-1.6
als 't goed is	5	18.57	32.41	-13.84

these categories were overrepresented (see Table 7) compared to the others, these results do confirm the intuition that there must be a relation between frequency of N-grams and syntactic constituency.
For both the selected 3-grams and 4-grams in Tables 8 and 9 t-tests revealed that the differences in percentage disagreement between the two context conditions are significant (for 3-grams: p=0.010 and for 4-grams p=0.030). Thus, it is safe to say that, on average, the pronunciation of words in the context of frequent N-grams differs more from the canonical form than the pronunciation of these words in arbitrary contexts. This finding also strongly suggests that many of the highly frequent N-grams in Tables 8 and 9 qualify for the status of MWE, if not for another reason, then at least because of their effect on pronunciation.

3.3.3 Qualitative analyses

In order to get more insight into the type of pronunciation variation that characterizes these 20 frequent 3- and 4-grams, the differences between the transcriptions in the two context conditions were also analyzed on a qualitative level based on the output of `Align`. In Table 10 we show how many of these

discrepancies were caused by deletions, substitutions and insertions.

Table 10
Average percentage disagreement (substitutions, deletions and insertions) for both context conditions.

av %	sub	del	ins	total
3-gram in MWE context	15.43	19.19	0.30	34.92
3-gram in other context	12.84	10.54	0.60	23.98
4-gram in MWE context	13.58	23.21	0.54	37.33
4-gram in other context	13.85	12.42	0.48	26.75

It is clear from this table that in both context conditions there are more deletions than insertions with respect to the canonical representations, which indicates that in both cases the actual pronunciations are reduced in comparison to their canonical reference. Since there are considerably more deletions in the condition "MWE context", it is legitimate to conclude that in this case the pronunciation of the individual words is more reduced than in the condition "other contexts". However, to get a better understanding of the type of reduction that affects the individual words when they appear in the context of N-grams, it is important to look not only at the number of deletions, but also at possible relations between deletions in individual words. Specifically, we are interested in the possibility that in "MWE context" the deletion of a cluster of phonemes occurs more often than in "other contexts". If deletion clusters are one of the specific phenomena for MWE contexts, they cannot be properly accounted for in the form of rewrite rules applied to individual words when generating a multi-pronunciation lexicon. To this end, we counted the number of deletion clusters of different length for all the words in the two context conditions (see Table 11).

Table 11
Distribution of deletion clusters of different sizes.

%	cluster 1	cluster 2	cluster 3	cluster 4
3-gram in MWE context	70.88	12.94	15.88	0.29
3-gram in other context	90.40	6.85	2.68	0.04
4-gram in MWE context	61.18	37.89	0.62	0.31
4-gram in other context	95.48	4.52	0.00	0.00

Table 11 clearly shows that the size and the distribution of deletion clusters are different in the two context conditions. In the condition "MWE context" there are clearly more deletion clusters of size 2, 3, and 4 than in the condition "other contexts". In other words, in the context of N-grams it is more common that sequences of two or three segments, therefore possibly whole syllables, are deleted. In addition, the fact that deletion clusters of a given size (i.e. 3 and 4 for 4-grams) are not found at all in the condition "other contexts" seems to suggest that there are pronunciation variants that are unique for the "MWE context" condition. Obviously, this is a point that deserves further investigation.

Qualitative analyses were also carried out for the data concerning the substitutions (cf. Table 12). In Table 10 we saw that the percentages of substitutions with respect to the canonical representation are similar in the two context conditions. Qualitative analyses of these substitutions also revealed that the processes underlying them are very similar. Table 12 shows that the most frequent substitutions concern processes such as voice assimilation and vowel reduction that are already known from the literature (Booij, 1995).

Table 12
Ten most frequent substitutions with percentage disagreement in both context conditions for 3-grams and 4- grams.

3-grams				4-grams			
MWE context		other context		MWE context		other context	
/t/-/d/	2.86	/t/-/d/	2.84	/t/-/d/	3.21	/t/-/d/	3.36
/k/-/g/	2.23	/d/-/t/	1.74	/k/-/g/	2.32	/k/-/g/	2.18
/v/-/f/	1.90	/k/-/g/	1.45	/v/-/f/	1.38	/d/-/t/	2.10
/E/-/@/	1.23	/s/-/z/	1.41	/A/-/@/	1.04	/A/-/@/	1.91
/I/-/@/	1.08	/A/-/@/	1.25	/d/-/t/	0.94	/I/-/@/	1.01
/d/-/t/	0.93	/v/-/f/	1.03	/E/-/@/	0.94	/s/-/z/	0.77
/a/-/@/	0.89	/I/-/@/	0.77	/p/-/b/	0.69	/z/-/s/	0.48
/a/-/A/	0.63	/a/-/A/	0.36	/s/-/z/	0.49	/v/-/f/	0.31
/f/-/v/	0.52	/a/-/@/	0.29	/n/-/N/	0.49	/n/-/m/	0.30
/z/-/s/	0.41	/E/-/@/	0.25	/e/-/@/	0.35	/A/-/a/	0.27
sum	12.68		11.40		11.85		12.69

4 Discussion

The analysis of frequent N-grams showed that a very large proportion (21%) of the words in the spontaneous speech in the CGN corpus are part of word sequences that occur frequently. This highly repetitive and predictable nature of extemporaneous speech deserves more attention in the future than it has received in the past. Furthermore, while compiling the set of frequent N-grams, we also found that there are quite a number of N-grams which occur frequently in very specific communicative settings and not at all in other settings. Whether this finding is coincidental or systematic can only be determined by comparing and analyzing more and larger spoken corpora than just the CGN.
In the CGN we have observed a tendency for frequent N-grams to consist of complete syntactic clauses, or at least opening part of a clause. Although this finding is intuitively plausible, we still need further research to understand its implications for psycholinguistics and speech technology.

The results presented in Section 3.3 clearly indicate that for all the words in the N-grams investigated the actual pronunciation is reduced with respect to its canonical representation. The amount of reduction in pronunciation is

mainly caused by the fact that many segments in the canonical representation appear to be deleted in the actual pronunciation. In addition, analyses of the substitutions observed reveal that many of these also concern reduction processes: i.e. substitutions of full vowels in the canonical transcriptions by schwas in the actual pronunciations. So, these results confirm those of previous investigations which have shown that in spontaneous casual speech words may be highly reduced (Ernestus et al., 2002; Keating, 1998; Kohler, 1990). However, in our study we wanted to determine whether this amount of reduction is characteristic of spontaneous speech across the board, or whether it is related to specific contexts, in particular those of frequent N-grams. To answer this question we examined the pronunciation variants of the same words in the context of N-grams and in all remaining contexts. The results of these analyses, presented in Sections 3.3.2 and 3.3.3, make it clear that for almost all the words investigated it holds that the degree of reduction is higher when these words appear in the context of frequent N-grams as opposed to when they appear in any other context. Moreover, analyses of the distribution of deletions reveal that in the context of frequent N-grams deletions tend to be more grouped together than in the other contexts, indicating that sometimes whole syllables are deleted in N-grams. Finally, the fact that the clustering pattern of deletions is different in the two context conditions and that certain cluster types are not found outside frequent N-grams indicates that 'MWE-like' N-grams probably contain unique pronunciation variants. These findings suggest that, at least for the purpose of pronunciation modelling, it is necessary to add a number of frequent N-grams with their characteristic pronunciation variants to the (pronunciation) lexicon. This may be a better solution than indiscriminate addition of all the pronunciation variants observed to the individual words in the lexicon, which, as shown in Kessens et al (1999), is counter-productive.

The most important reason to start the research reported in this paper was to determine whether these MWEs and their pronunciation variants require special handling in automatic speech recognition (ASR) and automatic phonetic transcription (APT). Previous research has shown that modelling pronunciation variation can be beneficial for both APT and ASR: for APT because the quality of the resulting transcriptions can be improved (Binnenpoorte et al., 2004a; Schiel, 1999); and for ASR, because the word error rates can be reduced (Strik and Cucchiarini, 1999). In ASR research it has also been shown that if too many variants are added, word error rates increase again. Specific modelling of pronunciation variation in MWEs has been studied in the field of ASR, but, as far as we know, not in the field of APT. In ASR, MWEs are referred to as phrases, word tuples, multiword units, or multiwords. Different criteria are used to select, usually a small number of, MWEs. Adding these MWEs and their pronunciation variants to the lexicon usually reduces word error rate. In general, the main goal of these studies is to reduce word error rate, and, consequently, no detailed study of pronunciation variation of

MWEs is carried out. In our study we did examine the type of pronunciation variation that characterizes a selected number of frequent MWEs and found that these exhibit uncommon pronunciation patterns that are not found in other contexts. We therefore suggest that these MWEs be included as lexical entries in the pronunciation lexicons employed in ASR and APT, because in both cases this is likely to improve the performance of the system.

5 Conclusions and perspectives for future research

In this paper we have presented an exploratory study of MWEs in spontaneous speech in which focusses on the pronunciation of MWEs in relation to ASR and APT. We have shown that the words composing the MWEs investigated do indeed exhibit different pronunciation patterns in the MWE context than in other contexts. This provides evidence for the fact that these MWEs require special treatment in ASR and APT.
The results of our study suggest that phonetically transcribed corpora are a valuable source for research into phenomena and problems that affect the performance of ASR and APT for conversational speech and that have so far been elusive. However, the practical problems encountered in this study also make it clear that eventually we will need phonetically transcribed corpora of unprecedented size. Therefore, it is essential to continue the research aimed at developing accurate automatic phonetic transcriptions of speech recordings. The results obtained with our medium size corpus already show a number of promising directions for that research.
Future research could also profit from the application of shallow syntactic parsing to the classification of N-grams that we have performed on the basis of the orthography alone. More detailed information about the type and the degree of completeness of the syntactic constituent formed by frequent N-grams should help in selecting the word sequences that are candidates for inclusion in a MWE lexicon.
Adding information about prosody, if only in the form of the strength of the juncture between adjacent words, is an obvious extension of the work reported in this paper. It seems evident that the presence of clear phonetic boundaries between adjacent words prevents the deletion of large phoneme clusters across the boundary. However, here too one will need large corpora with accurate transcriptions to support the research.

Acknowledgements

The authors would like to thank Nelleke Oostdijk, Peter-Arno Coppen and Bill Fletcher and the three anonymous reviewers for their careful and constructive

comments on earlier versions of this paper.

References

Beulen K, Ortmanns S, Eiden A, Martin S, Welling L, Overmann J, Ney H. Pronunciation modelling in the RWTH large vocabulary speech recognizer. In: *Proceedings of the ESCA Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*; 1998. p. 13-16.

Biber D, Johannson S, Leech G, Conrad S, Finegan E. *The Longman Grammar of Spoken and Written English*, Longman, Harlow, Essex, 1999. p. 987-1036.

Binnenpoorte D, Cucchiarini C, Strik H, Boves L. Improving Automatic Phonetic Transcription of Spontaneous Speech through Variant- Based Pronunciation Variation Modelling. In: *Proceedings of LREC 2004*, Lisbon; 2004. p. 681-684.

Binnenpoorte D, Cucchiarini C, Boves L, Strik H. Appendix to Multiword Expressions in Spoken Language: an exploratory study on pronunciation variation; 2004. http://lands.let.ru.nl/literature/mwe-appendix.html

Booij G. *The phonology of Dutch*, Clarendon Press, Oxford, 1995.

Cucchiarini, C. Assessing transcription agreement: methodological aspects. In: *Clinical Linguistics & Phonetics*, Vol. 10, No. 2; 1996. pp. 131-155.

Ernestus M, Baayen, H, Schreuder R. The Recognition of Reduced Word Forms. *Brain and Language*, 81; 2002. p. 162-173.

Finke M, Waibel A. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In: *Proceedings of EuroSpeech-97*, Rhodes; 1997. p. 2379-2382.

Hawkins, S. Roles and representations of systematic fine phonetic detail in speech understanding, *Journal of Phonetics*, Vol. 31; 2002, p. 373-405.

Keating P. Word-level phonetic variation in large speech corpora. In: A. Alexiadou, N. Fuhrhop, U. Kleinhenz, & P. Law (Eds.), *ZAS papers in linguistics*, vol. 11. Berlin: Zentrum für Allgemeine Sprachwissenschaft, Typologie und Universalienforschung; 1998. p. 35-50.

Kessens JM, Wester M, Strik H. Improving the performance of a Dutch CSR by modelling within-word and cross-word pronunciation variation. *Speech Communication*, 29 (2-4); 1999. p. 193-207.

Kessens JM, Cucchiarini C, Strik H. A data-driven method for modeling pronunciation variation. *Speech Communication*, 40 (4); 2003. p. 517-534.

Kohler KJ. Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In: W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling*, Dordrecht, Kluwer; 1990. p. 69-92.

Koster CHA. Transducing Text to Multiword Units. In: *Proceedings MEMURA 2004 workshop*, Lisbon; 2004. p. 31-38.

Nivre J, Nilsson J. Multiword Units in Syntactic Parsing. In: *Proceedings MEMURA 2004 workshop*, Lisbon; 2004. p. 39-46.

Nunberg G, Sag IA, Wasow T. Idioms, *Language*, 70; 1994. p. 491-538.

Odijk J. Reusable Lexical Representations for Idioms. In: *Proceedings LREC 2004*, Lisbon; 2004. p. 903-906.

Oostdijk NHJ. The design of the Spoken Dutch Corpus, In: P. Peters, P. Collins and A. Smith (Eds.): *New Frontiers of Corpus Research.* Amsterdam: Rodopi; 2002. p. 105-112.

Pallett, DS. A look at NIST's benchmark ASR tests: past, present, and future, In: *Proceedings Workshop Automatic Speech Recognition and Understanding*, 2003, p. 483-488.

Sag IA, Baldwin T, Bond F, Copestake A, Flickinger D. Multiword expressions: A pain in the neck for NLP. *LinGO Working Paper (2001-03)*; 2001. http://lingo.stanford.edu/pubs/WP2001-03.ps.gz.

Schiel F. Automatic Phonetic Transcription of Non-Prompted Speech. In: *Proceedings of the ICPhS 1999*, San Francisco; 1999. p. 607-610.

Sloboda T, Waibel A. Dictionary Learning for Spontaneous Speech Recognition. In: *Proceedings of ICSLP-96*, Philadelphia;1996. p. 2328-2331.

Strik H, Cucchiarini C. Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, Vol. 29 (2-4); 1999. p. 225-246.

Wong-Fillmore L. Individual Differences in Second Language Acquisition. In C. Fillmore, D. Kempler and W. Wang (Eds.) *Individual Differences in Language Ability and Language Behaviour.* Academic Press, New York; 1979. p. 203-228.

Yang Q, Martens J-P. On the importance of exception and cross-word rules for the data-driven creation of Lexica for ASR. In: *Proceedings 11th ProRisc Workshop*, Veldhoven, The Netherlands; 2000. p. 589-593.